DAOS Community Update

Johann Lombardi, TSC Chair
Per3S, May 28, 2024 in Paris

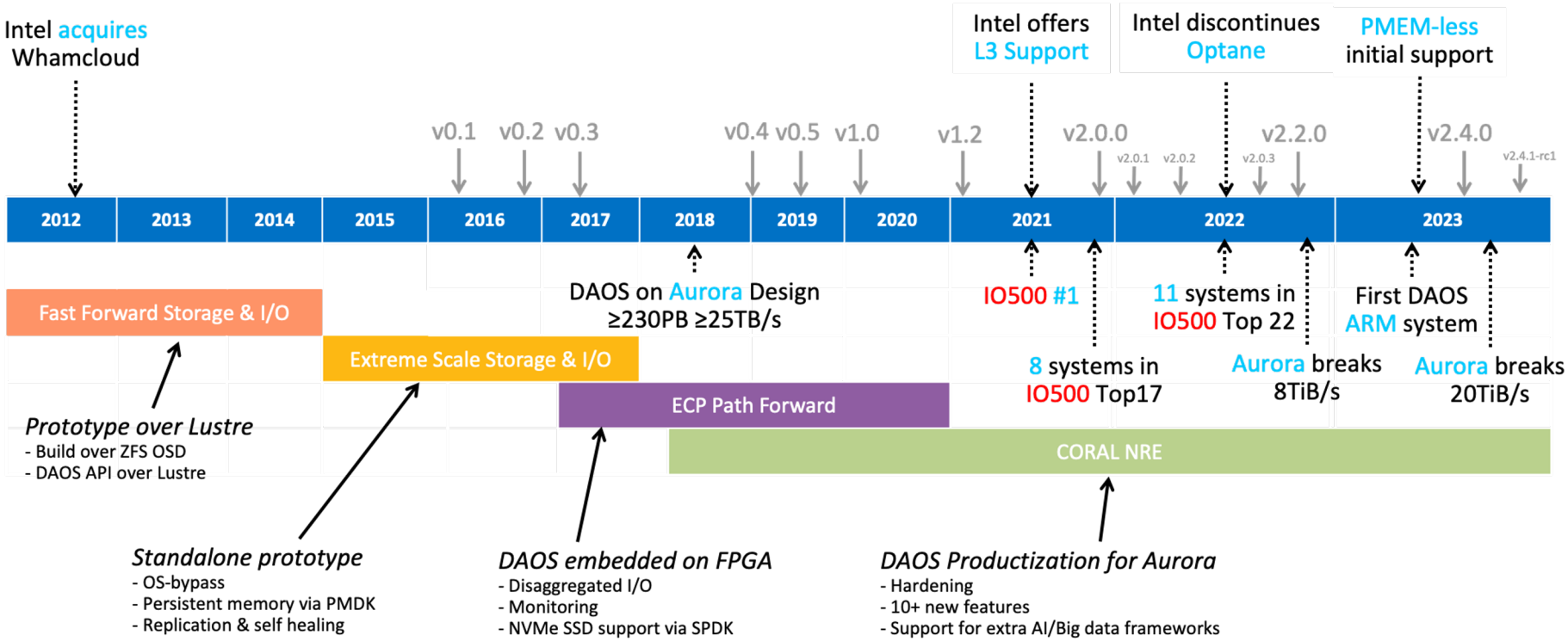
THE LINUX FOUNDATION

https://foundation.daos.io

daos

DAOS Introduction

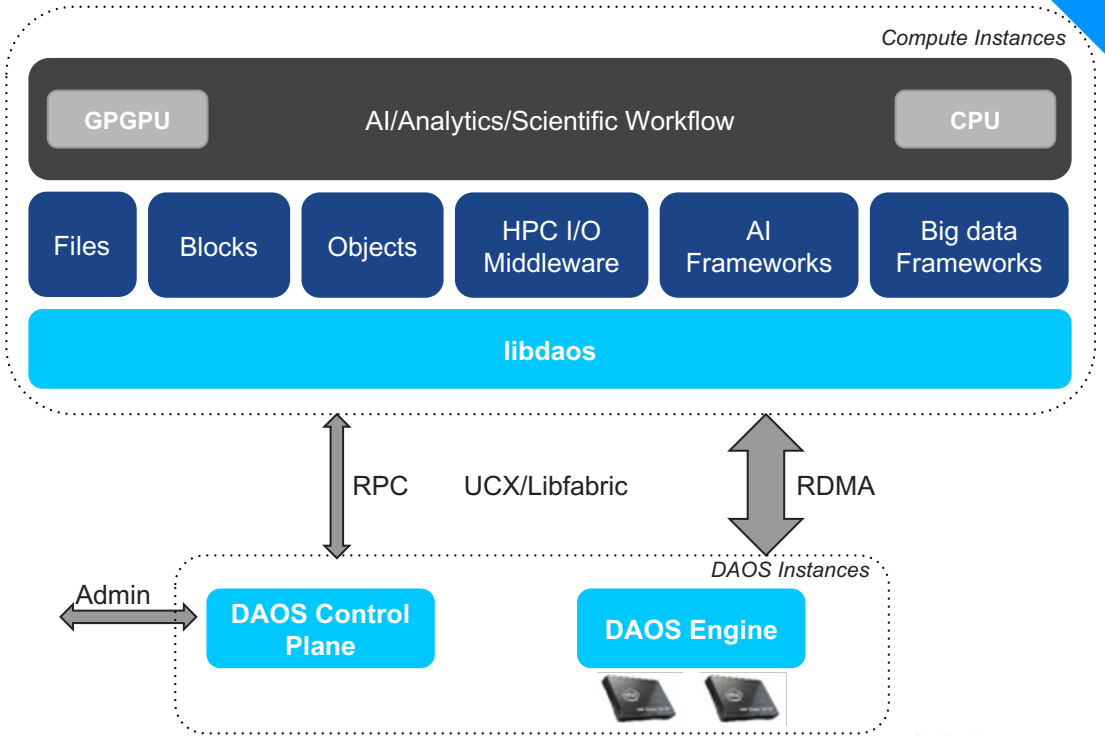
DAOS History



daos

DAOS: Nextgen Open Storage Platform

- Platform for innovation
- Files, blocks, objects and more
- Full end-to-end userspace
- Flexible built-in data protection
  - EC/replication with self-healing
- Flexible network layer
- Efficient single server
  - O(100)GB/s and O(1M) IOPS per server
- Highly scalable
  - TB/s and billions IOPS of aggregated performance
  - O(1M) client processes
- Time to first byte in O(10) µs



DAOS Design Fundamentals

- **No read-modify-write on I/O path (use versioning)**
- **No locking/DLM (use MVCC)**
- **No client tracking or client recovery**
- **No centralized (meta)data server**
- **No global object table**
- **Non-blocking I/O processing (futures & promises)**
- Serializable distributed transactions exposed to the users

Aurora DAOS System

- 1024x DAOS Storage nodes
  - 2x Xeon 5320 CPUs (ICX)
  - 512GB DRAM
  - 8TB Optane Persistent Memory 200
  - 244TB NVMe SSDs
  - 2x HPE Slingshot NICs
- Supported data protection schemes
  - No data protection
  - All EC flavors: 2+1, 2+2, 4+1, 4+2, 8+1, 8+2, 16+1 and 16+2
  - N-way replication
- Usable DAOS capacity
  - between 220PB and 249PB depending on redundancy level chosen

Aurora System Specifications

Compute Node 2 Intel Xeon scalable "Sapphire Rapids" processors; 6 Xe arch-based GPUs; Unified Memory Architecture; 8 fabric endpoints; RAMBO	**Software Stack** HPE Cray EX supercomputer software stack + Intel enhancements + data and learning	**GPU Architecture** Xe arch-based "Ponte Vecchio" GPU; Tile-based chiplets, HBM stack, Foveros 3D integration, 7nm
CPU-GPU Interconnect CPU-GPU: PCIe; GPU-GPU: Xe Link	**System Interconnect** Slingshot 11; Dragonfly topology with adaptive routing	**Network Switch** 25.6 Tb/s per switch, from 64–200 Gbs ports (25 GB/s per direction)
Peak Performance ≥ 2 Exaflop DP	**High-Performance Storage** ≥ 230 PB, ≥ 25 TB/s (DAOS)	**Programming Models** Intel oneAPI, MPI, OpenMP, C/C++, Fortran, SYCL/DPC++
Platform HPE Cray EX supercomputer	**Aggregate System Memory** > 10 PB	**Node Performance (TF)** > 130
System Size (# Nodes) > 9,000		

DAOS Performance - ISC'24 Production List

IO500

# ↑	INFORMATION						IO500			
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	
3	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	
4	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	
5	ISC24	Zuse Institute Berlin	Lise	Megware	DAOS	10	960	324.54	65.01	1,620.13	

IOR & FIND

EASY WRITE	20,693.63 GiB/s
EASY READ	12,122.87 GiB/s
HARD WRITE	4,216.34 GiB/s
HARD READ	9,706.55 GiB/s
FIND	229,672.10 kIOP/s

METADATA

EASY WRITE	60,985.13 kIOP/s
EASY STAT	225,295.35 kIOP/s
EASY DELETE	57,648.44 kIOP/s
HARD WRITE	33,827.19 kIOP/s
HARD READ	141,467.16 kIOP/s
HARD STAT	230,086.03 kIOP/s
HARD DELETE	62,196.78 kIOP/s

THE LINUX FOUNDATION

daos

Aurora IO500 Run

Features	Limits
Number of client nodes	512
Number of client endpoints	4k
Number of client processes	53k
Number of DAOS servers	642
Number of DAOS engines	1284
Largest Pool	160PiB
Largest file	8.5PiB
Total number of files	177 Billions
Number of files in a single directory	33 Billions

Foundation Overview

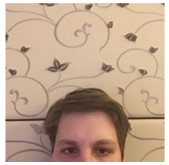
Who?

- DAOS Foundation formed by five organizations
  - Argonne National Laboratory
  - Enakta Labs
  - Google
  - Hewlett Packard Enterprise (HPE)
  - Intel Corporation
- Each organization had made some investment in DAOS and has been an early adopter
- Welcome any new members who wish to participate in DAOS development and direction (see how to join)

Board

Kevin Harms (Chair)
Argonne National Lab

Lance Evans
HPE



Denis Nuja (Treasurer)
Enakta Labs

Brad Hoagland
Intel Corporation

Dean Hildebrand
Google

Johann Lombardi
TSC Chair

THE LINUX FOUNDATION

daos

What?

- The DAOS Foundation exists to
  - Maintain DAOS as an open source project independent of any one organization
  - Foster the developer and user communities around DAOS
  - Guide the direction of the overall project

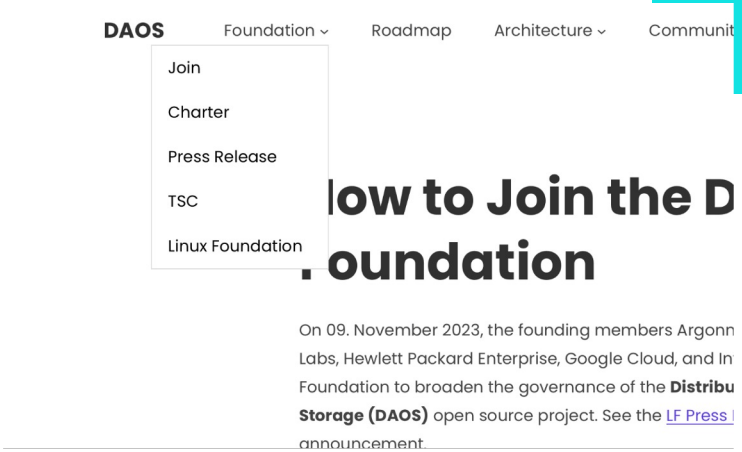
- Governing Board
  - Defines budget and approves expenses
  - Oversee efforts of other subcommittees
  - Approve roadmap provided by Technical Steering Committee (TSC)
  - Vote on matters as needed

Current Activities of Governing Board

- Establishing governing board procedures
- Complete LF legal updates for TSC Charter
- Transition elements to DAOS/LF ownership (drive, webhost, ...)
- Recruiting new members
- DUG planning
- Binary signing and hosting
- CI/CD

How? (to join)

- Two step process for any organization
  - Join the Linux Foundation (at any level)
  - Join the DAOS Foundation
- https://daos.io/?page_id=199
- DAOS Foundation
  - 3 levels with 5 fees



DAOS Foundation Membership Level	Annual Fees
Premier	25,000 USD
Premier for LF Associate Members	15,000 USD
General	15,000 USD
General for LF Associate Members	6,000 USD
Associate for LF Associate Members	0 USD

THE LINUX FOUNDATION

daos

Technical Steering Committee (1/2)

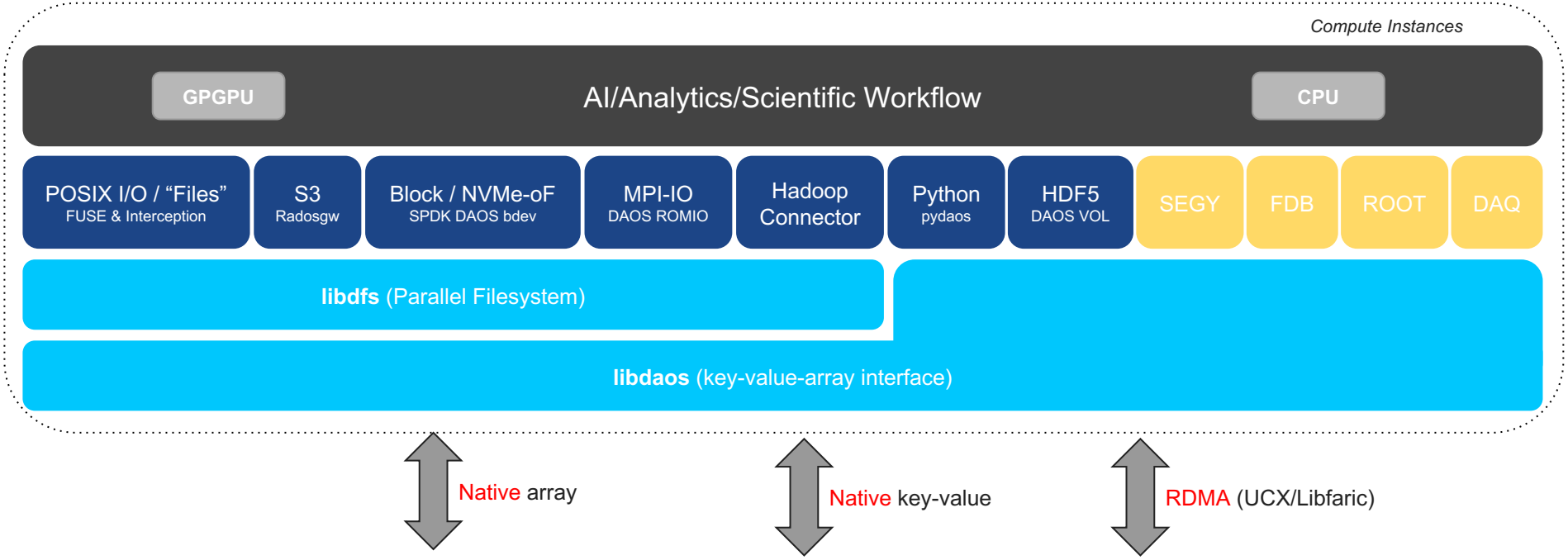
- Define community roadmap (2.8+)
  - Gather contributions from all community members
  - Train model
  - Publish roadmap on https://daos.io
  - What release should be LTS?
- Produce community releases (2.8+)
  - Track progress, review jira tickets & test results
  - Tag release and sign/distribute packages
  - Provide docker images
- Organize DAOS development
  - Simplify contributions
  - Organize gatekeeping (members, responsibilities, process)
  - Document contribution process

Technical Steering Committee (2/2)

- Community test infrastructure
  - Goal: artifacts and logs available to all contributors
  - Expand coverage
    - ARM/AMD
    - More fabrics
    - More linux distributions
    - Cloud environments
    - Focus on pmem-less mode
- Working groups
  - Open to anyone
  - Forums for DAOS users/administrators/contributors to exchange
  - Rotating schedule
  - 3 working groups with more to be added

DAOS Community Roadmap

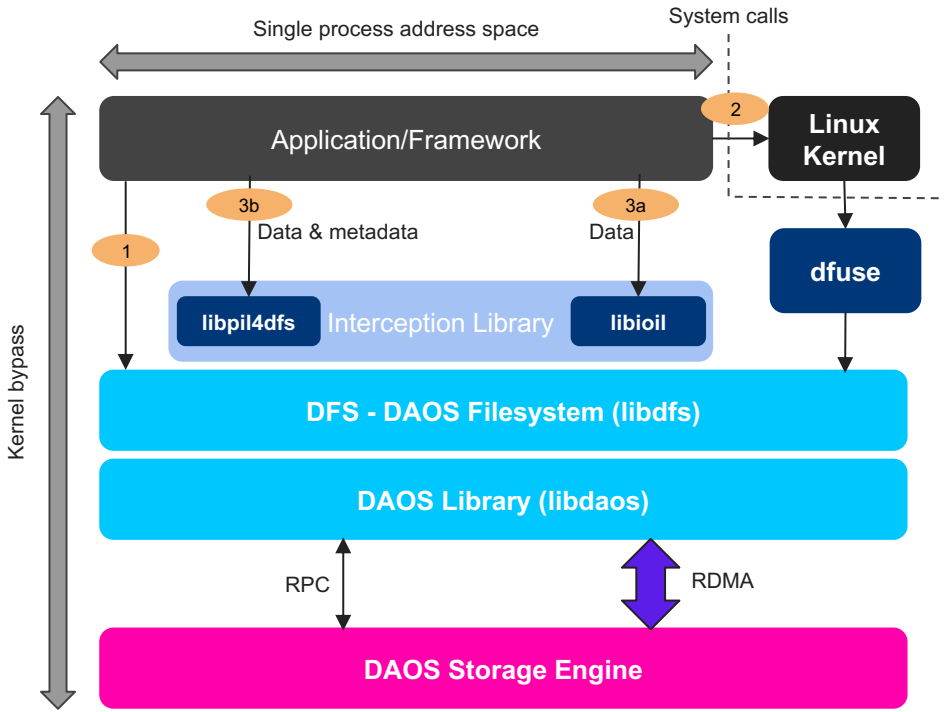
Software Ecosystem



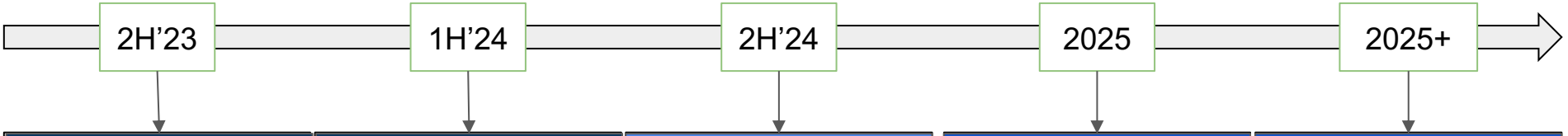
Generic I/O Middleware/frameworks

Domain-specific data models under development in co-design with partners

POSIX Support & Interception



1. Userspace DFS library with API like POSIX
   - **Require** application changes
   - Low latency & high concurrency
   - No caching
2. DFUSE daemon to support POSIX API
   - **No** application changes
   - VFS mount point & high latency
   - Caching by Linux kernel
3. DFUSE + Interception library
   - **No** application changes
   - 2 flavors using LD_PRELOAD
   - libioil
     - (f)read/write interception
     - Metadata via dfuse
   - libpil4dfs
     - Data & metadata interception
     - Aim at delivering same performance as #1 w/o any application change
     - Mmap & binary execution via fuse



DAOS 2.4
(Intel Release)

OS Packages:
- Leap 15.4/15.5
- RHEL/Rocky/Alma 8.6/8.8

Networking:
- HPE Slingshot support
- UCX Support
- Multiple engines per socket

Features:
- online drain/reintegration/add
- per-server fault domain
- EC parity rotation
- Dfuse readdir caching
- Streaming functions in libioil

Tech preview:
- pmem-less support phase 1
- pil4dfs

UX Improvements:
- LED management via Intel VMD

DAOS 2.6
(Intel Release)

OS Packages:
- Leap 15.5
- RHEL/Rocky/Alma 8.8/9.2

Networking:
- Change provider w/o reformat
- MD duplicate RPC detection

Features:
- pmem-less support phase 1
- libpil4dfs
- Intel VMD hotplug
- Delayed rebuild

Tech preview:
- Catastrophic recovery

UX Improvements:
- Improved version interoperability

DAOS 2.8
(DAOS Foundation Release)

OS Packages:
- Leap 15.6
- RHEL/Rocky/Alma 8.10/9.4
- Ubuntu 22.04

Networking:
- VNI support
- iWARP support

Features:
- Optimized object placement
- Mount POSIX snapshots RO
- Client telemetry

Tech preview:
- pmem-less support phase 2

UX Improvements:
- Reintegration of all pools
- Additional RAS events

DAOS 3.0
(DAOS Foundation Release)

OS Packages:
- Leap 15.6
- RHEL/Rocky/Alma 8.x
- Ubuntu 22.04
- RHEL/Rocky/Alma 9.x (+ARM64)

Networking:
- Multihomed network support

Features:
- pmem-less support phase 2
- WORM containers
- DSA optimizations
- Lustre integration
- Container parking/serialization
- Catastrophic recovery
- Fault domains beyond servers
- Engine scalability improvements

Tech preview:
- Rolling upgrade
- Incremental rebuild/reintegration
- Omni-Path Express support

DAOS 3.x
(DAOS Foundation Release)

OS Packages:
- Leap 15.7
- RHEL/Rocky/Alma 8.x
- Ubuntu 22.04
- RHEL/Rocky/Alma 9.x (+ARM64)

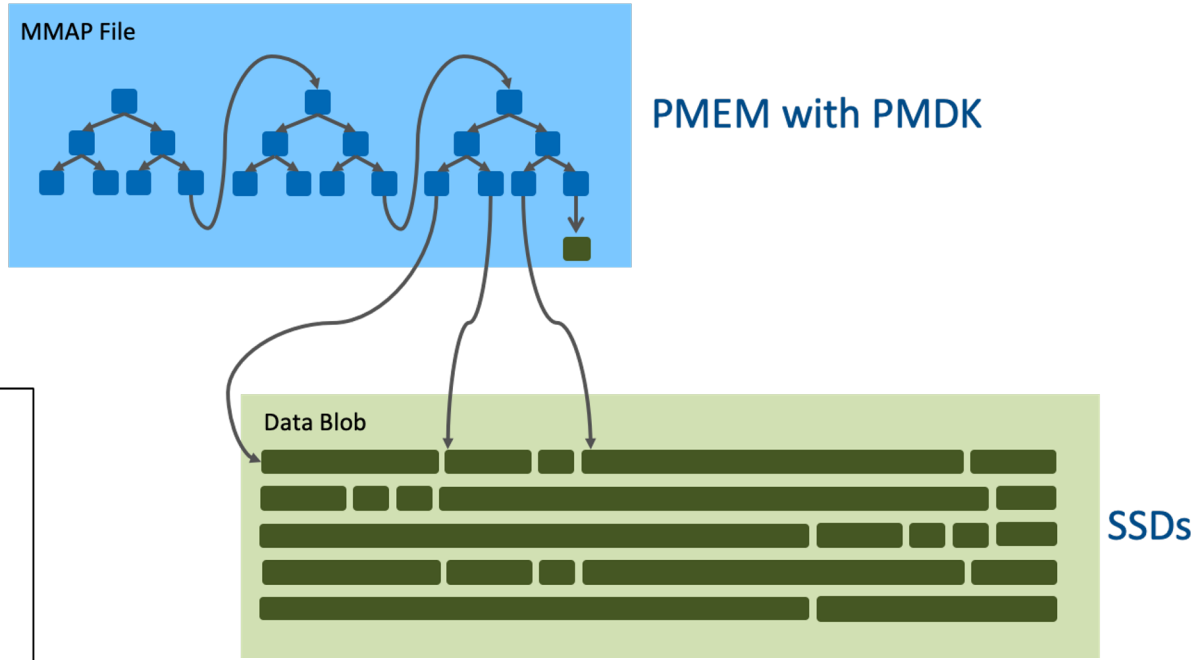
Features:
- Pool resizing
- Inline compression (QAT)
- Inline encryption (QAT)
- Inline deduplication
- Middleware catastrophic recovery
- Progressive layout
- Pipeline API
- SQL support with predicate pushdown
- Distributed transactions
- Pool/container freeze
- CXL SSD support / QLC
- Tiered container phase 1

THE LINUX FOUNDATION

Color coding schema:
- Committed (or released) release/features
- In-planning release/features
- Future possible release/features

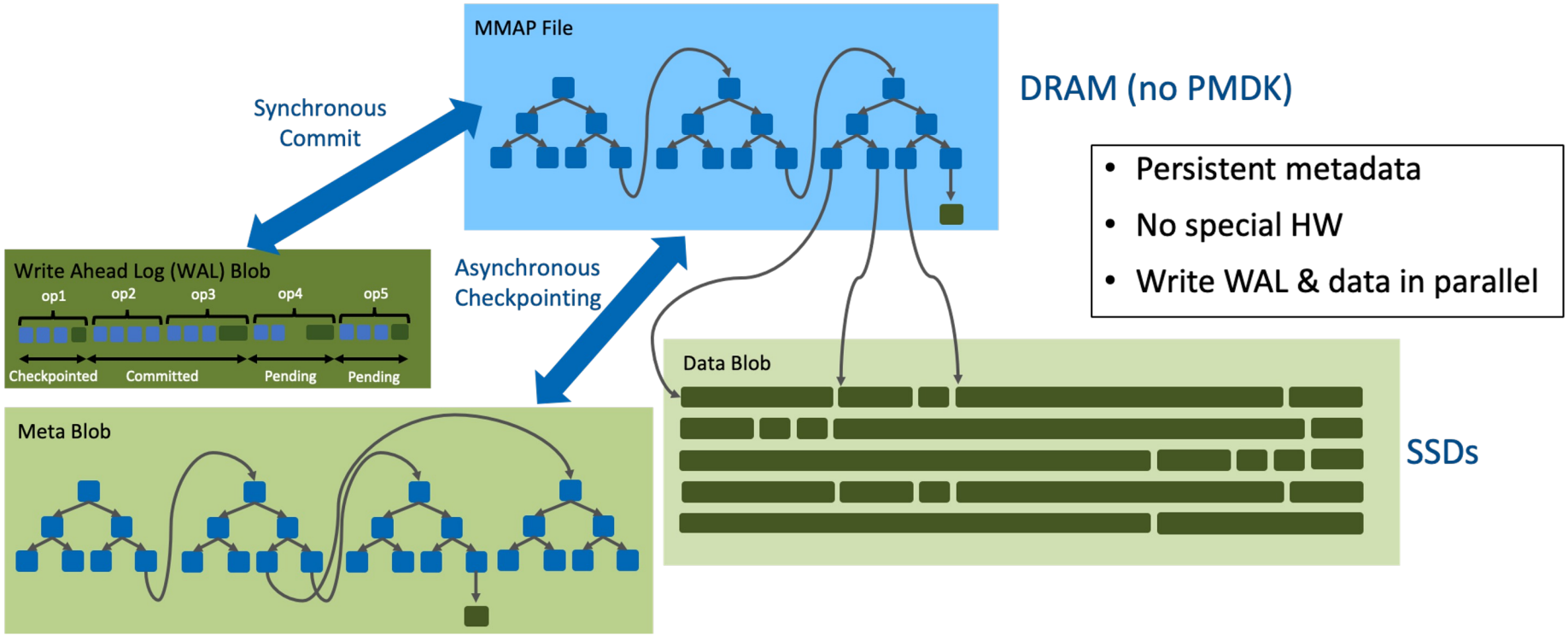
daos

Pmem Mode



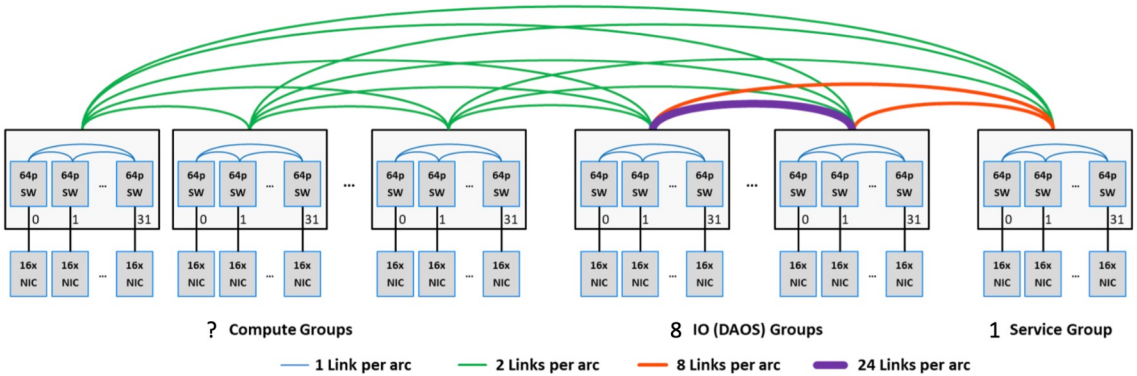
- Persistent metadata
- Require Intel Optane PMEM (or NVDIMM-N)
- App Direct mode
- Mode used on Aurora

Pmem-less Mode



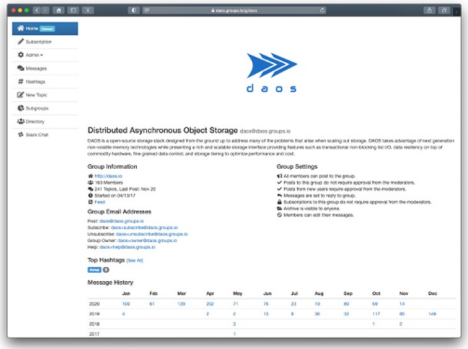
Optimized Object Placement

- Placement aware of the network topology
  - E.g. DragonFly fabric
- Performance domains complementary to fault domains
- Different strategies
  - Spread redundancy groups as widely as possible across different performance domains
  - as possible

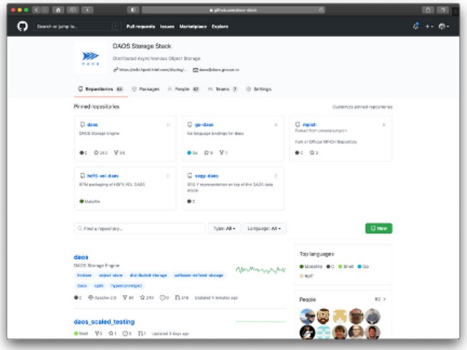


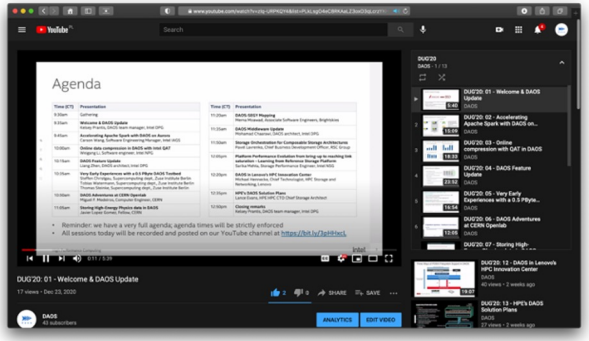
Aurora Network Topology
(from Kevin Harms' presentation at DUG'22)

Resources



- Foundation website: https://daos.io/
- Github: https://github.com/daos-stack/daos
- Online doc: https://docs.daos.io
- Mailing list & slack: https://daos.groups.io
- YouTube channel: http://video.daos.io
- 8th DAOS User Group (DUG'24) at SC'24 in Atlanta





THE LINUX FOUNDATION

daos

Thanks

DAOS Foundation Levels

- **Premier Membership**
  - Each Premier Member can appoint a voting member to the DAOS Foundation's Governing Board, its Outreach Committee, and to any other committee that the DAOS Foundation may establish (including the TSC).
- **General Membership**
  - The group of all General Members annually elect up to three voting representatives to the DAOS Foundation's Governing Board (depending on the number of General Members).
  - Each General Member can appoint a non-voting member to the DAOS Foundation's Outreach Committee.
- **Associate Membership**
  - The Associate Members can participate in the activities of the DAOS Foundation, but have no seat on the Governing Board and no voting rights.